

Hydration of a B-DNA Fragment in the Method of Atom–atom Correlation Functions with the Reference Interaction Site Model Approximation

D.A. Tikhonov, R.V. Polozov,

Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, Pushchino, Moscow Region, 142292, Russia

E.G. Timoshenko *, Yu.A. Kuznetsov, A.V. Gorelov, K.A. Dawson

Irish Centre for Colloid Science and Biomaterials, Department of Chemistry, University College Dublin, Belfield, Dublin 4, Ireland

(February 1, 2008)

We propose an efficient numerical algorithm for solving integral equations of the theory of liquids in the Reference Interaction Site Model (RISM) approximation for infinitely dilute solution of macromolecules with a large number of atoms. The algorithm is based on applying the nonstationary iterative methods for solving systems of linear algebraic equations. We calculate the solvent–solute atom–atom correlation functions for a fragment of the B-DNA duplex d(GGGGG)-d(CCCCC) in infinitely dilute aqueous solution. The obtained results are compared with available experimental data and results from computer simulations.

I. INTRODUCTION

Interactions of nucleic acids with proteins, lipids and other ligands in many respects are determined by the solvation properties of these biomolecules. The structure of the DNA molecule itself in the cytoplasmic environment of a cell is strongly dependent on its hydration. Information about the structure of water shell around many biomolecules is available primarily from X-ray and neutron scattering experiments¹ on their crystalline forms, as well as from NMR studies² of such molecules in solution.

There is a large number of reviews devoted to the problem of biopolymer hydration³. One of common approaches to these difficult problems is based on computer simulations using either Molecular Dynamics or Monte Carlo methods^{4–6}. The main limitation of these direct simulation techniques is in the huge computational expense for reaching the equilibrium state even for comparatively small molecular weights. A promising alternative for addressing the equilibrium issues is based on the methods of statistical mechanics and, particularly, on the method of integral equations of the theory of liquids commonly known as RISM^{7–9}. The main advantage of the method based on the Reference Interaction Site Model (RISM) equations is that here one explicitly deals with all quantities averaged over the equilibrium Gibbs distribution. This results in a considerable reduction of computational expenses for many problems compared to those of direct simulation methods. An example of application of the RISM technique to hydration of biomolecules is the calculation of the alanine dipeptide carried out in Ref. 10.

Practically, in the limit of infinite dilution the RISM method requires numerical solution of a system of integral equations for the solute–solvent pair correlation functions of order proportional to the number of atoms in a macromolecule. Here the solvent–solvent correlation functions can be found beforehand and the conformation of a macromolecule is assumed fixed. We note that memory requirements for keeping the matrix of unknown variables could be quite considerable — a few terabytes of RAM or more even for a DNA duplex of only 5 bases.

A number of coarse–grained models^{11,12} for complicated macromolecules have been suggested to eliminate this difficulty in the RISM method. For instance, in Refs. 11 the solvation of model DNA structures has been investigated using such coarse–graining with each nucleotide pair being approximated by a potential centre and the solvent by a binary mixture of charged particles. Similar coarse–grained models may be of interest for addressing non–specific hydration effects, which are fairly independent of the detailed molecular structure and thus universal for a given class of biomolecules. In Ref. 13 the hydration of a linear nonpolar chain in aqueous solution has been considered to elucidate the general behaviour of hydration of long macromolecular chains. Unfortunately, errors resulting from a coarse–graining procedure are practically impossible to estimate and therefore such models seem to be too remote from the complexities of the real DNA molecule.

Here we suggest an efficient numerical algorithm for solving the RISM equations that makes it feasible to study a detailed model for short duplexes of DNA in water in a quite moderate computational time, even on a PC and,

*Corresponding author; E-mail: Edward.Timoshenko@ucd.ie

most importantly, avoiding the necessity to deal explicitly with huge matrices. The main idea of this algorithm is to use the Newton–Raphson scheme with a specially approximated Jacobi matrix with consequent application of the nonstationary iterative methods for solving large systems of linear equations. A review of modern developments in the latter field, on which our approach heavily relies, may be found in Ref. 14.

In this work we describe the new algorithm for calculation of the solvent–solute correlation functions of a macromolecule in aqueous solution. This is then applied to study of the hydration of B-DNA duplexes.

II. BASIC EQUATIONS OF THE RISM THEORY

Various integral equation techniques have proved to be successful for a wide range of applications from simple monoatomic liquids to complex multicomponent mixtures of molecular liquids. Generally, the correlation functions satisfy the infinite chain of Bogoliubov–Born–Green–Kirkwood–Yvon (BBGKY) integro-differential equations. To achieve a practical progress one possibility is to apply a closure relation expressing high-order correlation functions in terms of a few low-order ones. Many different closure relations have been suggested, each having its own region of validity, which can be established by comparing the results of calculations with available data from experiment or computer simulations.

Another particularly fruitful strategy is based on the Ornstein–Zernike (OZ) equation, which relates the direct, $\tilde{C}(k)$, and the total, $\tilde{h}(k)$, correlation functions in the Fourier space,

$$(1 - \rho\tilde{C}(k))(1 + \rho\tilde{h}(k)) = 1, \quad (1)$$

where ρ is the density of the liquid. Here the original and the Fourier conjugated functions, the latter being designated by the tilde, are related as follows,

$$\tilde{x}(k) = \frac{4\pi}{k} \int_0^\infty r dr \sin(kr) x(r), \quad x(r) = \int_0^\infty k dk \frac{\sin(kr)}{2\pi^2 r} \tilde{x}(k). \quad (2)$$

The total correlation function is defined in such a way that $h(r) + 1$ is the probability density to find an atom at the distance r from another one. Eq. (1) is an exact integral equation between $h(r)$ and the direct correlation function, $C(r)$, which can be established based on the diagrammatic cluster expansion. This has a simple physical meaning. $C(R)$ is said to be caused by the two-body interaction potential between a pair of atoms, $u(r)$, whereas the indirect correlation function, $\gamma(r) = h(r) - C(r)$, contains contributions of all other atoms.

The direct correlation function also has a diagrammatic representation, which can be viewed as the exact closure relation between the two unknown functions. However, one would like to express such a relation in an explicit form which is popularly written as,

$$C(r) = \exp(-u(r)/k_B T + \gamma(r) + \mathcal{B}[\gamma(r)]) - \gamma(r) - 1, \quad (3)$$

where k_B is the Boltzmann constant, T is the absolute temperature and $\mathcal{B}[\gamma(r)]$ is called the bridge functional. Unfortunately, no strict analytical expression can be found for the latter.

Therefore, in practice, one has to revert to some kind of an approximate closure relation inspired by partial resummations of the diagrammatic expansion⁹. Let us mention here two of the most widely used relations,

$$\mathcal{B}[\gamma(r)] = 0, \quad (4)$$

$$\mathcal{B}[\gamma(r)] = \ln(1 + \gamma(r)) - \gamma(r), \quad (5)$$

which are called the hyper-netted chain (HNC) and the Percus–Yevick (PY) closures respectively. It is well established that, generally, the approximation (4) is more accurate for long-range potentials, whilst approximation (5) works better for short-range potentials.

Generalisation of the OZ equation to multicomponent mixtures of molecular liquids seems to be relatively straightforward. The conformation of a molecule is assumed to be fixed, so that the correlation function between the i -th and j -th sites (atoms) inside the same molecule are given by $v_{ij}(r) = \delta(r - r_{ij})/4\pi r^2$, where r_{ij} are the known interatomic distances. In the Fourier space the analogue of Eq. (1) can be written in the matrix form,

$$\left(\hat{v}(k) + \hat{\rho}\hat{h}(k) \right) \left(\hat{v}(k)^{-1} - \hat{\rho}\hat{c}(k) \right) = \hat{1}, \quad (6)$$

where \hat{v} is a block diagonal matrix with the number of blocks equal to the number of distinct molecules and the elements of each block are equal to $v_{ij}(k) = \sin(kr_{ij})/kr_{ij}$. The matrix $\hat{\rho}$ contains the partial atomic densities of the

mixture and possesses the same block structure as the matrix \hat{v} . The form (6) is often called the Site-Site Ornstein-Zernike equation (SSOZ), and jointly with a particular closure relation (3) such as e.g. HNC or PY is loosely referred to as the RISM theory. We shall favour this terminology for historical reasons despite it may not be the most logical one⁹.

The important point is, however, that the relation of these equations to the diagrammatic expansion is no longer preserved in the site-site version of the formalism. So, in a sense, the SSOZ equation may be simply viewed as a formal definition of the site-site direct correlation function⁸, with no particular deep meaning attached to the analogues of HNC or PY closures. The shortcomings of the RISM technique, that are possibly related to this circumstance, are well known and include the problems with the long-ranged orientational correlations and the equation of state (to discussion of these problems and attempts to improve the RISM theory we refer the reader to standard reviews such as e.g. 9). Despite these deficiencies, we believe that the RISM remains a very useful tool for studying hydration of complex biomolecules¹⁵, where no other method can really compare to it in computational power. However, based on the standard techniques in Ref.¹⁵ it was possible to consider the system of 413 atoms only thanks to using a powerful supercomputer.

For a two-component molecular solution let us denote by ρ the density of the solvent and by ρ_u the density of the solute. In the limit of infinite dilution, $\rho_u/\rho \rightarrow 0$, Eq. (6) is decoupled,

$$\hat{\gamma}^v = \hat{v}^v \hat{c}^v \hat{v}^v (\hat{1} - \rho \hat{c}^v \hat{v}^v)^{-1} - \hat{c}^v(k), \quad (7)$$

$$\hat{\gamma}^{uv} = \hat{v}^u \hat{c}^{uv} (\hat{v}^v + \rho \hat{h}^v) - \hat{c}^{uv}, \quad (8)$$

where the superscripts v and u refer to the solvent and the solute respectively.

If we are interested in hydration of a N -atomic molecule, Eq. (8) for the solute-solvent correlation functions of all atoms in the molecule with the O and H atoms of water may be rewritten in the Fourier space as follows,

$$G = (\hat{V} \otimes \hat{W} - \hat{1}_{2N}) C, \quad (9)$$

where G and C are vectors of size $2N$ containing the correlation functions,

$$G^T = [\tilde{\gamma}_{1H}(k), \dots, \tilde{\gamma}_{NH}(k), \tilde{\gamma}_{1O}(k), \dots, \tilde{\gamma}_{NO}(k)], \quad (10)$$

$$C^T = [\tilde{c}_{1H}(k), \dots, \tilde{c}_{NH}(k), \tilde{c}_{1O}(k), \dots, \tilde{c}_{NO}(k)], \quad (11)$$

where the superscript T stands for the transposition. Here the first subscript of a correlation function is the atom number in the solute molecule and the second superscript designates the atom type in the water molecule. The structure matrix, \hat{W} , of size N is expressed via the interatomic distances, r_{ij} ,

$$W_{ij} = \frac{\sin(kr_{ij})}{kr_{ij}}. \quad (12)$$

The matrix \hat{V} expresses the geometrical and correlational structure of the water,

$$\hat{V} = \begin{bmatrix} v_1 & v_2 & v_3 \\ 2v_3 & v_4 \end{bmatrix}, \quad (13)$$

where its components are,

$$\begin{aligned} v_1 &= 1 + \rho \tilde{h}_{HH}(k), & v_2 &= \sin(k r_{HH})/k r_{HH} + \rho \tilde{h}_{HH}(k), \\ v_3 &= \sin(k r_{OH})/k r_{OH} + \rho \tilde{h}_{OH}(k), & v_4 &= 1 + \rho \tilde{h}_{OO}(k). \end{aligned} \quad (14)$$

Here r_{HH} and r_{OH} are the interatomic distances in the water molecule, ρ is the density of water and \tilde{h}_{HH} , \tilde{h}_{OH} and \tilde{h}_{OO} are the total solvent-solvent correlation functions which may be found beforehand.

To avoid divergences we apply the HNC closure relation (4) using the renormalization procedure of the Coulomb potential proposed by Ng^{16,9},

$$c_{iX}^s(r) = \exp(-U_{iX}(r)/k_B T + f_{iX}(r) + \gamma_{iX}^s(r)) - \gamma_{iX}^s(r) - 1, \quad (15)$$

where $U_{iX}(r)$ is the non-Coulombic part of the interatomic potential, the subscript i is the atom number in the solute molecule and index X refers to atoms O or H in the water molecule. Here the function f_{iX} is the renormalized Coulomb potential expressed through the charges q_i and q_X :

$$f_{iX}(r) = -q_i q_X \frac{1 - \text{erf}(r)}{r k_B T}. \quad (16)$$

The short-range parts of the correlation functions, which are designated by the superscript s above, are related to the complete correlation functions as follows,

$$\tilde{\gamma}_{iX}(k) = \tilde{\gamma}_{iX}^s(k) - \tilde{f}_{iX}(k), \quad \tilde{c}_{iX}(k) = \tilde{c}_{iX}^s(k) + \tilde{f}_{iX}(k). \quad (17)$$

We also note that the introduction of the short-range correlation functions allows one to perform the Fourier transformation using the fast Fourier transform technique.

III. THE NUMERICAL ALGORITHM

In this section we propose an algorithm for numerical solution of the system of $2N$ integral equation (9, 15) for large values of N . We assume that the solvent density, ρ , the temperature $k_B T$, the matrices \hat{W} and \hat{V} and the interaction potentials appearing in the closure relation (15) are given.

First of all, let us introduce discretisation by choosing L grid points uniformly distributed in the r - and k -spaces,

$$\Delta r \Delta k = \frac{\pi}{L}, \quad r_i = i \Delta r, \quad k_j = j \Delta k, \quad (i, j = 1 \dots L). \quad (18)$$

We denote by \vec{X} the $2NL$ -component vector, which approximates the set of functions, $\tilde{\gamma}_{iX}^s(k)$ in $j \Delta k$ grid points,

$$X^T = [\tilde{\gamma}_{1H}^s(\Delta k), \dots, \tilde{\gamma}_{1H}^s(L \Delta k), \dots, \tilde{\gamma}_{NO}^s(L \Delta k)]. \quad (19)$$

Then we introduce the discrete Fourier transformation operators \mathcal{F}^f and \mathcal{F}^b ,

$$\hat{F}^f = \frac{4\pi(\Delta r)^2}{\Delta k} \hat{S}, \quad \hat{F}^b = \frac{(\Delta k)^2}{2\pi^2 \Delta r} \hat{S}, \quad S_{ij} = \frac{j}{i} \sin\left(\frac{\pi}{L} ij\right), \quad (i, j = 1, \dots, L). \quad (20)$$

Eq. (9) becomes a system of $2NL$ nonlinear algebraic equations,

$$\vec{X} - \vec{Z}[\vec{X}] \equiv \vec{X} - \hat{A}(\hat{1}_{2N} \otimes \hat{F}^f)\vec{C}[(\hat{1}_{2N} \otimes \hat{F}^b)\vec{X}] - \vec{Q} = 0. \quad (21)$$

Here \hat{A} is a matrix of size $2NL$ consisting of diagonal blocks of size L , which is the discrete analogue of the matrix $\hat{V} \otimes \hat{W} - \hat{1}_{2N}$, and $\vec{Q} = (\hat{V} \otimes \hat{W})[\tilde{f}_{1H}(k), \dots, \tilde{f}_{NO}(k)]^T$. For discretisation of the closure relation (15) let us also introduce the function $\vec{C}[\vec{Y}]$,

$$C_l \equiv C_l[Y_l] = M_l \exp(Y_l) - Y_l - 1, \quad (l = 1, \dots, 2NL), \quad (22)$$

where M_l are elements of a $2NL$ -dimensional vector \vec{M} ,

$$M^T = [\exp(f_{1H}(r) - U_{1H}(r)), \dots, \exp(f_{NO}(r) - U_{NO}(r))]. \quad (23)$$

The simplest algorithm for solving the system of nonlinear algebraic equations (21) would be the Picard method of direct iterations. Thus, given an initial approximation, $\vec{X}^{(0)}$, the improved solution is found from the recurrent relation,

$$\vec{X}^{(n+1)} = \vec{Z}[\vec{X}^{(n)}], \quad (n = 0, 1, \dots). \quad (24)$$

The iterations should stop when the norm of $\vec{X}^{(n+1)} - \vec{X}^{(n)}$ becomes smaller than a predefined small value. Some modifications of the method of direct iterations are often utilized, which involve several previous iterations for providing a better convergence. A good example of such modifications, based on a vector extrapolation, was used in Ref. 17. The direct iteration method for solving of the RISM equations was used in Ref. 18, where the principle of “monotonic discrepancy” was applied to improve convergence. Unfortunately, in practice the convergence of the direct iteration method seems to be an exception rather than the rule.

The Newton–Raphson algorithm appears to be much more efficient for solving equations (21). Here iterations are performed by the following scheme,

$$\vec{X}^{(n+1)} = \vec{X}^{(n)} + \Delta \vec{X}^{(n)}, \quad (25)$$

$$\Delta \vec{X}^{(n)} = \hat{J}^{-1} (\vec{Z}[\vec{X}^{(n)}] - \vec{X}^{(n)}), \quad (26)$$

where \hat{J} is the Jacobi matrix obtained by differentiating Eq. (21):

$$\hat{J} \equiv \hat{1}_{2NL} - \frac{d\vec{Z}}{d\vec{X}} \equiv \hat{1}_{2NL} - \frac{2}{L} \hat{A} (\hat{1}_{2N} \otimes \hat{S}) \hat{D} (\hat{1}_{2N} \otimes \hat{S}). \quad (27)$$

The elements of the diagonal matrix \hat{D} are given by,

$$D_{ll} = M_l \exp(Y_l) - 1, \quad (l = 1, \dots, 2NL). \quad (28)$$

The disadvantage of the Newton–Raphson method is that at each iteration step it is necessary to solve a system of linear equations (26) of size $2NL$. It is a rather large system, since, even in the case of comparatively small N , the number of grid points L should be of order of thousands to achieve a good discretization. Despite that the convergence of the Newton–Raphson iterations is rather fast, in practice the expenses for solving the system of linear equations with the Jacobi matrix of size $2NL$ are extremely high.

Gillan¹⁹ has suggested a more efficient algorithm which is essentially a hybrid of the Newton–Raphson and the Picard schemes for the ‘coarse’ and ‘fine’ parts of the solution respectively, the former being obtained as an expansion in the basis of so-called roof functions. In Ref. 20 it was noted that the expansion of the ‘coarse’ part as a truncated Fourier series yields even a better convergence. Thus, the main feature of both the Gillan and the LMV methods is that the solution is sought in a cycle of the Newton–Raphson steps for a small-size system of equations for the expansion coefficients of the ‘coarse’ part, and a consequent Picard refinement for the solution. We note, however, that this attractive scheme²¹ is difficult to apply in practice for large molecules. Indeed, if the size of the system of equations for the ‘coarse’ part is taken relatively small, the consequent direct Picard iteration would not converge. If, on the contrary, this size is taken sufficiently large, the corresponding Jacobi matrix would not fit into the computer memory.

Here propose a somewhat different strategy which could be viewed as the Newton–Raphson scheme with an approximated Jacobi matrix. Namely, let us choose a subset of the grid points in the coordinate and k -spaces,

$$r'_i = i m \Delta r, \quad k'_j = j \Delta k, \quad (i, j = 1 \dots \frac{L}{m}), \quad (29)$$

where m is an integer number. The Jacobi matrix calculated on this subset is denoted as \hat{J}_m and its size is $2NL/m$. If we denote as k'' the complementary subset of components, we suggest to perform iterations according to the following formula,

$$\vec{R}^{(n)} = \vec{X}^{(n)} - \vec{Z}[\vec{X}^{(n)}], \quad (30)$$

$$\vec{X}^{(n+1)}(k') = \vec{X}^{(n)}(k') - \hat{J}_m^{-1} \vec{R}^{(n)}(k'), \quad (31)$$

$$\vec{X}^{(n+1)}(k'') = \vec{X}^{(n)}(k'') - \vec{R}^{(n)}(k''). \quad (32)$$

Thus, the approximated inverse Jacobi matrix has the elements of the exact inverse to the reduced Jacobi matrix for the k' subset indices, and the unity matrix for the k'' subset indices. Again, iterations are carried out until the norm of the vector $\vec{R}^{(n)}$ becomes smaller than a predefined small value.

The main feature of this particular scheme, which is an alternative to the Gillan and LMW methods, is that, in a sense, it combines the Newton–Rapson and Picard iterative schemes performed simultaneously. The reduction parameter m , which could be chosen as an integer power of 2, can be decreased during iterations if the convergence is slowing down, or can be increased if the convergence is improving. Practically, to achieve a good initial convergence, the size of the reduced Jacobi matrix $2NL/m$ is still going to be rather large for the molecule we are going to study here. Fortunately, now this difficulty can be eliminated.

For calculation of the inverse matrix in Eq. (31), i.e. solving a system of linear equations, we propose to apply the nonstationary iterative methods. Basically, these methods are versions of the conjugated gradients method. The main advantage of these methods is a better convergence for a wider class of matrices. Also, compared to the Gauss method, they do not require to keep the matrix of the system in the computer memory. Instead, it would suffice here to provide a procedure of matrix–vector multiplication, $\vec{y} = \hat{J}\vec{x}$, for arbitrary vector \vec{x} . If the number of grid points L is equal to an integer power of 2, then one can use the fast Fourier transform technique for calculation of vector–matrix multiplications. It is straightforward to see that this procedure requires LN^2 floating point operations for large N . The additional calculational efforts per iteration in the nonstationary iteration methods are proportional to NL . For example, the *Bi*–CGSTAB method¹⁴ requires two calculations of vector–matrix products, and an additional $48NL$ floating point operations. Overall, the memory requirements here are restricted to $20NL$ floating point units.

IV. NUMERICAL RESULTS

In this section we study fragments of the double helix DNA formed by nucleotide pairs G:C. We have calculated fragments of different lengths (from 1 to 5 pairs), so that the number of atoms N varied from 63 to 315.

We assumed the geometry of these fragments to be standard²² and the interaction potentials of atoms in different pairs in a fragment with water equal. According to Ref. 23 the non-Coulombic part of the interaction potentials of DNA atoms with water atoms can be taken in the form,

$$U_{iX}(r) = A_{iX}/r^{12} - B_{iX}/r^6. \quad (33)$$

If an atom forms a hydrogen bond we use another expression for the potential,

$$U'_{iX}(r) = A'_{iX}/r^{12} - B'_{iX}/r^{10}. \quad (34)$$

The latter form is used in the AMBER force field potential parametrisation²⁴. It is added to merely fine-tune the distances between atoms forming hydrogen bonds, which, in principle, are well enough represented by the Coulombic attraction alone. The potential parameters were taken from Ref. 23 and the atomic charges were taken from Ref. 25. As solvent we use water at the normal density 1 g/sm³ and the temperature 25°C. Correlation functions of pure water were calculated using Eq. (7) in the HNC closure for the potential model TIPS with the parameters taken from Ref. 26. The charge of the oxygen atom in the water molecule we consider to be equal to $q_0 = -0.86$ and that of the hydrogen atom — to $q_H = -q_0/2$. The interatomic distances in the DNA fragment were approximated up to the precision of 0.01 Å to reduce the number of distinct elements in the matrix \hat{W} . This allowed us to strongly reduce the requirements for RAM. Nevertheless, our calculations show that the correlation functions change rather insignificantly due to this round-off. The number of grid points and step size were taken equal to $L = 1024$ and $\Delta r = 0.025$ Å. The reduction parameter was set to $m = 8$, so that the size of the linear system is equal to 80640. The *TFQMR* nonstationary method¹⁴ was applied for solving the reduced system of linear equations. The precision parameters for solution of the nonlinear equations, ϵ , and of the linear ones, δ , were equal to 10^{-5} and 10^{-6} respectively. Calculations were carried out as follows. First, we calculate the correlation functions of one nucleotide pair. The output of this calculation is then used as input for calculation of two nucleotide pairs, and so on.

In Tab. I we present the computational expense of the problem. One can see that the computational time does not grow catastrophically with the increasing system size. The number of iterations, N^{ITER} , remains sufficiently low for fragments lengths considered here. This is because the calculational time is determined mainly by the time required for matrix–vector multiplications and it grows approximately as the second power of the reduced linear system size, $2NL/m$.

In Fig. 1 we exhibit the calculated correlation functions, $h(r) + 1$, for some of the DNA duplex atoms with atoms of water. Heights of the first maximum of the correlation functions vary from 4 to 0.4. High peaks of correlation functions are observed for heteroatoms in bases and for the oxygen atoms in the sugar–phosphate backbone. Heights of the first peak, which are less than unity, correspond to those atoms which are buried in the DNA molecule and thus screened by other atoms from a direct contact with water molecules.

Now, let us consider separately the hydration of bases, the sugar–phosphate backbone and also the characteristic features of hydration of the major and minor grooves of DNA.

A. Hydration of bases

The highest correlation peaks correspond to the nitrogen atoms of guanine N2 (see Fig. 1a) and N7 (Fig. 1b). The position of the peak in the correlation function between the hydrogen of water molecules and N7 in guanine is shifted to the left compared to the peak of the correlation between the oxygen of water and N7. The positions of the peaks in correlations atom N2 — oxygen in water and atom N2 — hydrogen in water are nearly identical. Also, the height of the first peak in correlations of atom N2 — hydrogen is much smaller than that for the atom N7. According to the results from Monte Carlo simulations⁶ this difference in behavior and degree of hydration of N7 and N2 is a consequence of that N7 atom is a proton acceptor, while N2 is a proton donor in the hydrogen bond. Obviously this results in different water molecule orientations around these atoms, and this is manifested in the considered correlation functions. The characteristic difference in the hydration of these atoms along the duplex chain is that the hydration of N7 does not change along the chain, whilst the hydration of N2 increases from ends to the centre approximately in 1.5 times.

The oxygen O2 in cytosine (see Fig. 1c) participates in a hydrogen bond with an amino group of guanine through the atom 1H2. The oxygen O6 in guanine (Fig. 1d) forms a hydrogen bond with an amino group of cytosine through

atom 2H4. The correlation functions of these two atoms are similar to each other. The characteristic feature of these functions is increasing of the first peak heights from ends to the centre analogously to N2 atom in guanine. Similarly, the oxygen atoms participate in forming hydrogen bonds with water as can be seen from shifted peaks in their correlation functions with the hydrogen of water molecules.

The hydrogen atoms in amino groups of guanine, 2H2 (Fig. 1e) and 1H2 (Fig. 1f), and 1H4 (Fig. 1g) and 2H4 (Fig. 1h) of cytosine differ significantly in their hydration behaviour. Protons participating in hydrogen bonds with oxygens O2 and O6 produce approximately twice lower height of the first correlation peak compared to that of the correlation functions for 2H2 of guanine and 1H4 of cytosine. This is not surprising since hydrogens 2H2 and 1H4 are open for water molecules, while 2H4 and 1H2 atoms hydrate weakly.

Hydration mechanisms of N1 atom in guanine (see Fig. 1i) and N3 (Fig. 1j) atom in cytosine, which both participate in hydrogen bonds between guanine and cytosine, are very different from each other. The distinction can be observed in the height of the first correlation peak with the hydrogen atom. N3 atom in cytosine possesses a large partial charge (approximately -0.7), resulting in a shift of the correlation peak with the hydrogen and in its comparatively large height. At the same time, the first peak in the correlation function with the oxygen is somewhat lower than unity for inner nucleotides of the DNA duplex. This means that, as in the case of the N1 atom, water weakly penetrates into the inter-planar space of neighbouring nucleotide pairs.

N3 atom of guanine (Fig. 1k) is partially screened from water molecules by amino group atoms. Thus, it hydrates in a much smaller degree than N7 atom (Fig. 1b), though atom N3 possesses a larger partial charge (-0.63) compared to the N7 atom (-0.54). N9 nitrogen atom in guanine (Fig. 1l) and N1 atom in cytosine (Fig. 1m) do not hydrate at all since they form a glycosidic bond.

B. Hydration of the sugar–phosphate backbone

Correlation functions of O1P atom of C residue on the 5' end are smaller than that in the other four nucleotide pairs (Fig. 1n). On the contrary, the O2P atom of C-nucleotide on 5' end possesses higher correlation peak compared to other pairs (Fig. 1o). We can observe the same picture for oxygen atoms in the phosphate group of G-nucleotide residue (Fig. 1p and Fig. 1q).

The phosphorus atoms in C and G residues hydrate nearly equally, except for the 5' end atoms, where the first correlation peak becomes broader (Fig. 1r and Fig. 1s). As for the hydration of deoxyribose, we should note that the ring atom O4* hydrates most strongly, the correlation peak on the 5' end being smaller for the G residue (Fig. 1t and Fig. 1u). O3* oxygens in G and C residues do not hydrate except for atoms on 3' ends, which can be seen from Figs. 1v, 1w. Hydration of G O5* (Fig. 1x) and C O5* (Fig. 1y) atoms does not differ from each other. Unlike the O3* atom (Fig. 1v, 1w), the hydration of end- and center- O5* atoms is the same. The rest of carbon and hydrogen atoms in the sugar–phosphate backbone hydrate rather weakly.

C. Hydration of the major and the minor grooves

Strong hydration observed for N7 (Fig. 1b) and O6 (Fig. 1d) atoms of guanine and N4 atom of cytosine (Fig. 1z) qualitatively corresponds to the hydration regions W1 and W2 presented in Ref. 27, while hydration of O2 atom in cytosine (Fig. 1c) and N3 (Fig. 1k), N2 (Fig. 1a), 2H2 (Fig. 1e) atoms in guanine correspond to the region S1.

On the other hand, as we have already mentioned, the first peaks in the correlation functions grow from ends of the duplex to its center for heteroatoms N2 (Fig. 1a), O6 (Fig. 1d), N4 (Fig. 1z) and O2 (Fig. 1c) participating in hydrogen bonds between guanine and cytosine. On the other hand, it is known from the calculations of hydration for monoatomic nonpolar chains¹³ and can be explained from the steric considerations, that the first peak of correlation function is much higher for end-atoms than for center-atoms. Such distinction in hydration behavior may be explained by two factors: first, by the presence of large partial charges on heteroatoms, and also by the geometrical structure of a Watson–Crick pair G:C. Presumably, this effect has a cooperative origin and is connected to the orientational ordering of water molecules in the grooves.

The values of correlation functions calculated by us are in good agreement with the results of Ref. 28, where the microscopic densities of oxygen and hydrogen atoms of water were calculated in the region contiguous to the major and the minor grooves of DNA using the mean-force potential formalism.

V. CONCLUSION AND DISCUSSIONS

In the current work we have developed an efficient algorithm for solving the integral equations of the theory of liquids in the RISM approximation for studying hydration of macromolecules. The current treatment allows one to study hydration without applying any additional simplifying assumptions about the molecular structure of a macromolecule. The algorithm can be used to consider comparatively long (several dozens of nucleotide pairs) biologically functional DNA fragments using small computers. It can also be used for studying hydration of promoters and genome terminators.

In particular, here we have studied the hydration of B-DNA fragments consisting of identical nucleotide pairs. Our results lead us to the conclusion that the hydration of nucleotide atoms in the DNA chain depends mainly only on the structure of its neighbouring environment, i.e. it is local by its nature. Our conclusions may as well be extended to an arbitrary sequence of G:C pairs. One expects that the hydration of the sugar–phosphate backbone depends very weakly on the particular sequence of base pairs. However, study of the dependence of the hydration of nucleotide pairs in the chain on the type of neighbouring pairs is beyond the scope of current paper.

In our calculations we assumed a rigid geometrical structure of DNA duplex. There are reasons to believe that the hydration of a flexible or a bent double helix differs from that of a rod-like structure. We believe that the flexibility of five nucleotide pairs is small enough for these effects to appear. In principle, the RISM method allows one to calculate the mean–force potentials as function of a particular molecular geometry, which would finally permit study of the problem of the optimal DNA geometry in water by Monte Carlo method¹⁸.

The reason we chose the length of the DNA duplex equal to five nucleotide pairs is that, as our calculations show, for this length the influence of the ends on the hydration of the central pair is very small. This can be clearly seen from the data, since in most cases the correlation functions of the penultimate and the central pairs are nearly identical. Thus, one can assume that the hydration of the central nucleotide pair in a short chain differs insignificantly from that of a nucleotide pair inside an infinitely long chain. We note that in this respect the solvation of monoatomic nonpolar chains in the aqueous environment is analogous to that of DNA¹³.

We realize that the current molecular model is limited in many respects since the DNA molecule was electro-neutral here due to a parametric account of the phosphate screening. Nevertheless, the existing experience of calculations of DNA hydration by various methods indicates that this approximation appears to be quite adequate in the region restricted by the first coordination sphere^{5,6,29}. The method used by us can be extended to include free mobile ions in the model and thus to study macromolecules in ionic solutions.

ACKNOWLEDGMENTS

The authors thank Professor V.I. Poltev and Dr O.A. Mornev for interesting discussions and useful suggestions and Dr S.S. Kolesnikov and Dr Yu.V. Bobkov for technical assistance. We also acknowledge the support of the Centre for High Performance Computing Applications, University College Dublin, Ireland. This work was supported in part by grant INTAS-93-2084.

¹ H. Savage, A. Wlodawer. *Meth. Enzymol.* **127**, 162 (1986); R.E. Dickerson. *Meth. Enzymol.* **211**, 67 (1992).

² G. Otting, E. Liepinsh, K. Wuthrich. *Science* **254**, 974 (1991); E. Liepinsh, G. Otting, K. Wuthrich. *Nucl. Acids Res.* **20**, 6549 (1992); M.G. Kubinec, D.E. Wemmer. *J. Amer. Chem. Soc.* **114**, 8739 (1992).

³ J. Texter. *Prog. Biophys. Mol. Biol.* **33**, 83 (1978); H.J.C. Berendsen, J.P.M. Postma, W.F. Van Gunsteren, J. Hermans. *Proc. 14th Jerusalem Symp. on Quantum Chemistry and Biochemistry*, (B. Pullman Ed., Reidel, Dordrecht, The Netherlands, 331, 1987); W. Saenger. *Annu. Rev. Biophys. Biophys. Chem.* **16**, 93 (1987); E. Westhof, D.L. Beveridge. "Water Science Reviews", (F. Franks Ed., Cambridge University Press. 1990. **5**, 24, 1990); E. Westhof. "Structural Water of Nuclear Acids". (E. Westhof Ed., Academic Press. 11, 1990); H.M. Berman. *Curr. Opin. Struct. Biol.* **1**, 423 (1991).

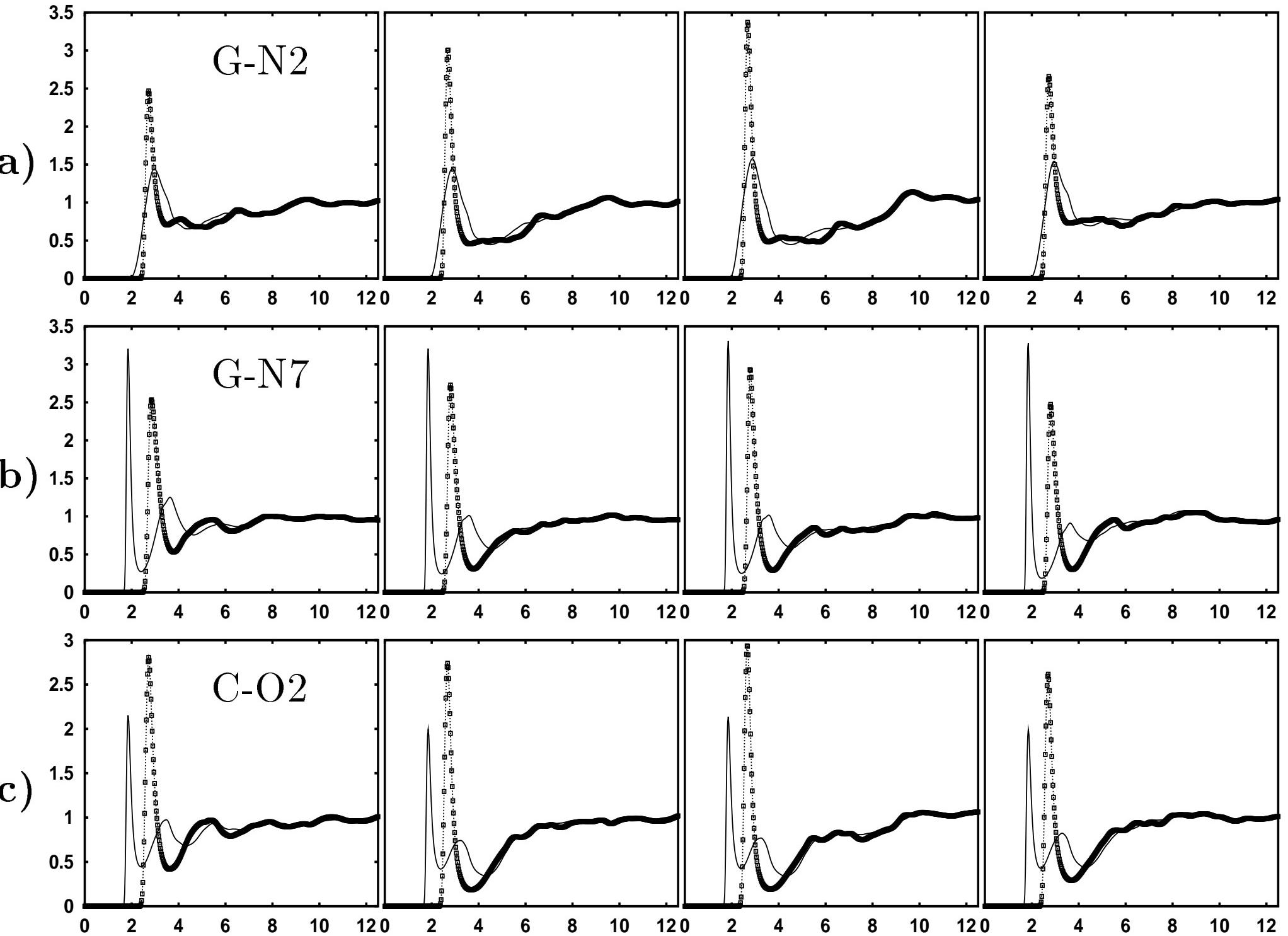
⁴ E. Clementi, G. Corongiu. *Biopolymers*. **20**, 551 (1981); E. Clementi, G. Corongiu. *Biopolymers*. **20**, 2427 (1981); E. Clementi, G. Corongiu. *Biopolymers*. **21**, 763 (1982); R. Bacquet, P.J. Rossky. *J. Phys. Chem.* **88**, 2660 (1984); C.S. Murthy, R.J. Bacquet, P.J. Rossky. *J. Phys. Chem.* **89**, 701 (1985); B.E. Hingerty, R.H. Ritche, T.L. Ferrel, J.E. Turner. *Biopolymers* **24**, 427 (1985); P. Mills, C.F. Anderson, T.M. Record (Jr). *J. Phys. Chem.* **89**, 3984 (1985); W.F. Van Gunsteren, H.J.C. Berendsen, R.G. Geurtsen, H.R.J. Zwinderman. *Ann. N.Y. Acad. Sci.* **482**, 297 (1986); R.J. Elliot, J.M. Goodfellow. *J. Theor. Biol.* **127**, 403 (1987). F. Eisenhaber, V.G. Tumanyan, F. Eisenmenger, W. Gunia. *Biopolymers*. **28**, 741 (1989). B. Jayaram,

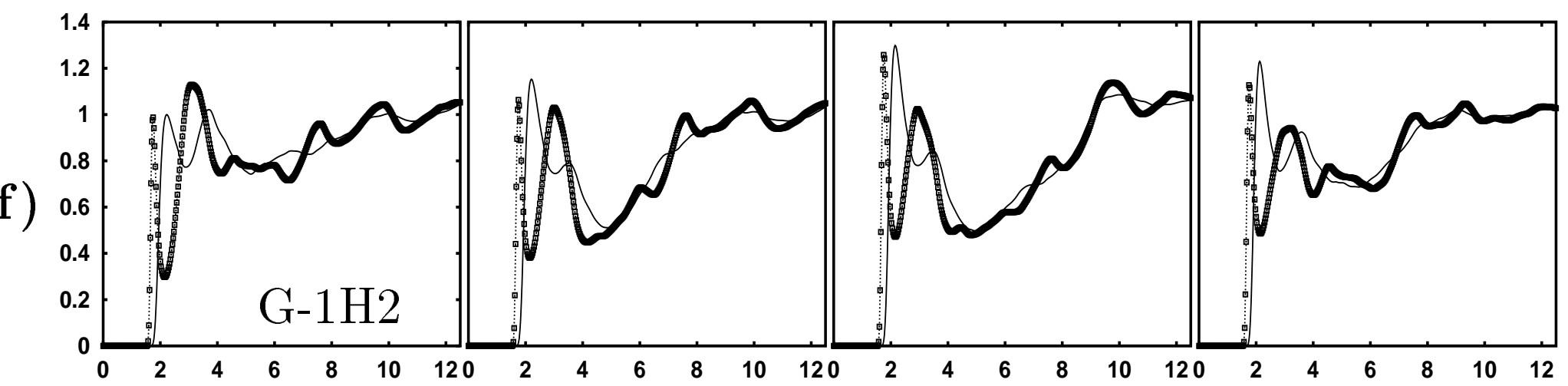
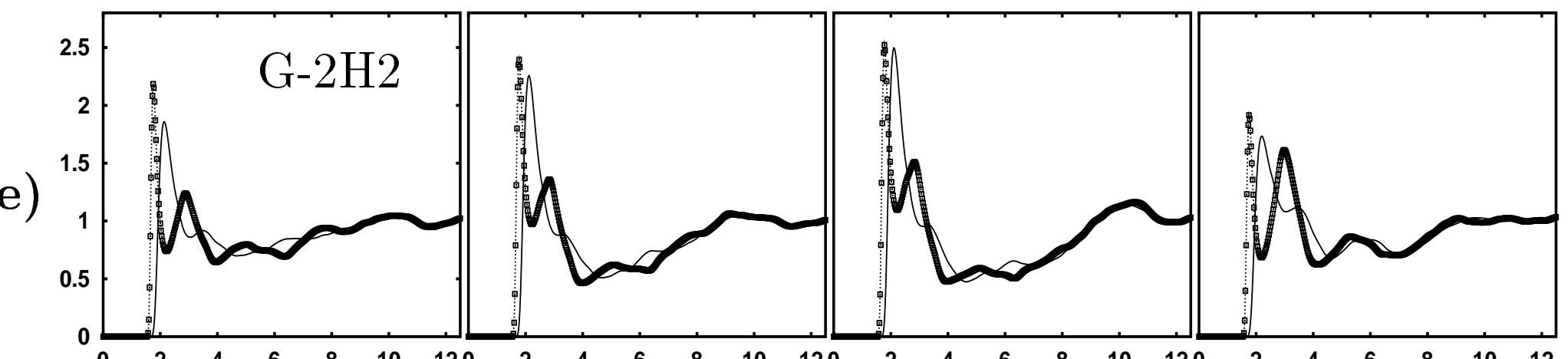
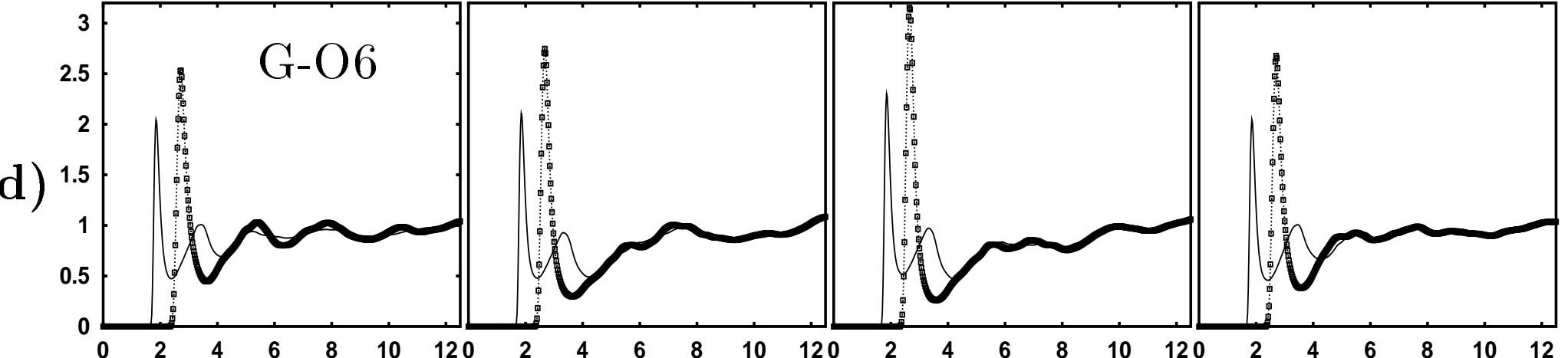
- S. Swaminathan, D.L. Beveridge. *Macromolecules*. **23**, 3156 (1990). T.R. Forester, I.R. McDonald. *Mol. Phys.* 1991. V. 72. P. 643. P. Herzyk, J.M. Goodfellow, S. Neidle. *J. Biomol. Struct. Dyn.* **9**, 363 (1991). D.M. York, T.A. Darden, L.G. Pedersen. *J. Chem. Phys.* **99**, 8345 (1993). D.L. Beveridge, G. Ravisbunker. *Curr. Opin. Struct. Biol.* **4**, 246 (1994). T.E. Cheatham, J.L. Miller, T. Fox, T.A. Darden, P.A. Kollman. *P.A. J. Amer. Chem. Soc.* **117**, 4193 (1995); T.E. Cheatham, P.A. Kollman. *J. Mol. Biol.* **259**, 434 (1996).
- ⁵ V.P. Chuprina, U. Heinemann, A.A. Nurislamov, P. Zielenkiewicz, R.E. Dickerson, W. Saenger. *Proc. Natl. Acad. Sci. U.S.A.* **88**, 593 (1991).
- ⁶ A.V. Teplukhin, V.I. Poltev, V.P. Chuprina. *Biopolymers*. **32**, 1445 (1992).
- ⁷ D. Chandler, H.C. Andersen. *J. Chem. Phys.* **57**(5), 1930 (1972).
- ⁸ G. Stell, G.N. Patey, J.S. Hoye, *Adv. Chem. Phys.* **48**, 183 (1981); P.T. Cummings, G. Stell, *Mol. Phys.* **46**, 383 (1982).
- ⁹ P.A. Monson, G.P. Morris. *Adv. Chem. Phys.* **77**, 451 (1990).
- ¹⁰ B.M. Pettitt, M. Karplus, P.J. Rossky. *J. Phys. Chem.* **90**(23), 6335 (1986).
- ¹¹ F. Hirata, P.J. Rossky, B.M. Pettitt. *J. Chem. Phys.* **78**, 4133 (1983); F. Hirata, R.M. Levy. *Chem. Phys. Lett.* **136**, 267 (1987); F. Hirata, R.M. Levy. *J. Phys. Chem.* **93**, 479 (1989).
- ¹² B. Svensson, C.E. Woodward, *J. Phys. Chem.* **99**, 1614 (1995).
- ¹³ D.A. Tikhonov, G.N. Sarkisov. *Zhurn. Fiz. Himii*. **71**, 480 (1997) (in Russian).
- ¹⁴ C.T. Kelley. "Iterative Methods for Linear and Nonlinear Equations", (SIAM Frontiers 16, 1995).
- ¹⁵ A. Kitao, F. Hirata, N. Go, *J. Phys. Chem.* **97**, 10231 (1993).
- ¹⁶ K.-Ch. Ng. *J. Chem. Phys.* **61**, 2680 (1974).
- ¹⁷ H.H. Homeier, S. Rast, H. Krienke. *Comp. Phys. Commun.* **92**, 188 (1995).
- ¹⁸ S.K. Talitskiih, P.K. Halatur. *Zhurn. Fiz. Himii* **69**, 5 (1997) (in Russian).
- ¹⁹ M.J. Gillan, *Mol. Phys.* **38**, 1781 (1979).
- ²⁰ S. Labik, A. Malijevsky, P. Vonka, *Mol. Phys.* **56**, 709 (1985).
- ²¹ T. Ichiye, D.J. Haymet, *J. Chem. Phys.* **89**, 4315 (1988).
- ²² Landoldt-Bornstein. "Numerical Data and Functional Relationships in Science and Technology", (New Series, W. Saenger, ed., Springer-Verlag, Berlin, 1989).
- ²³ V.I. Poltev, A.V. Teplukhin, G.G. Malenkov. *Int. J. Quant. Chem.* **42**, 1499 (1992).
- ²⁴ S.J. Weiner, P.A. Kollman, D.T. Nygen, D.A. Case. *J. Comp. Chem.* **7**, 230 (1986).
- ²⁵ V.B. Zhurkin, V.I. Poltev, V.L. Florent'ev. *Mol. Biol.* **14**, 1116 (1980) (in Russian).
- ²⁶ W.L. Jorgensen. *J. Amer. Chem. Soc.* **103**, 341 (1981).
- ²⁷ B. Schneider, D.M. Cohen, L. Schleifer, A.R. Srinivasan, W.K. Olson, H.M. Berman. *Biophys. J.* **65**, 2291 (1993).
- ²⁸ G. Hummer, D.M. Soumpasis. "Proc. Eighth Conversations in the Discipline Biomolecular Stereodynamics. (R.H. Sarma, M.H. Sarma, Eds. Adenine Press, Schenectady, **2**, 273, 1994); G. Hummer, D.M. Soumpasis. *Phys. Rev. E* **50**, 5085 (1994); G. Hummer, D.M. Soumpasis, A.E. Garcia. *Biophys. J.* **66**, A25 (1994).
- ²⁹ G. Hummer, A.E. Garcia, D.M. Soumpasis. *Biophys. J.* **68**, 1639 (1995).

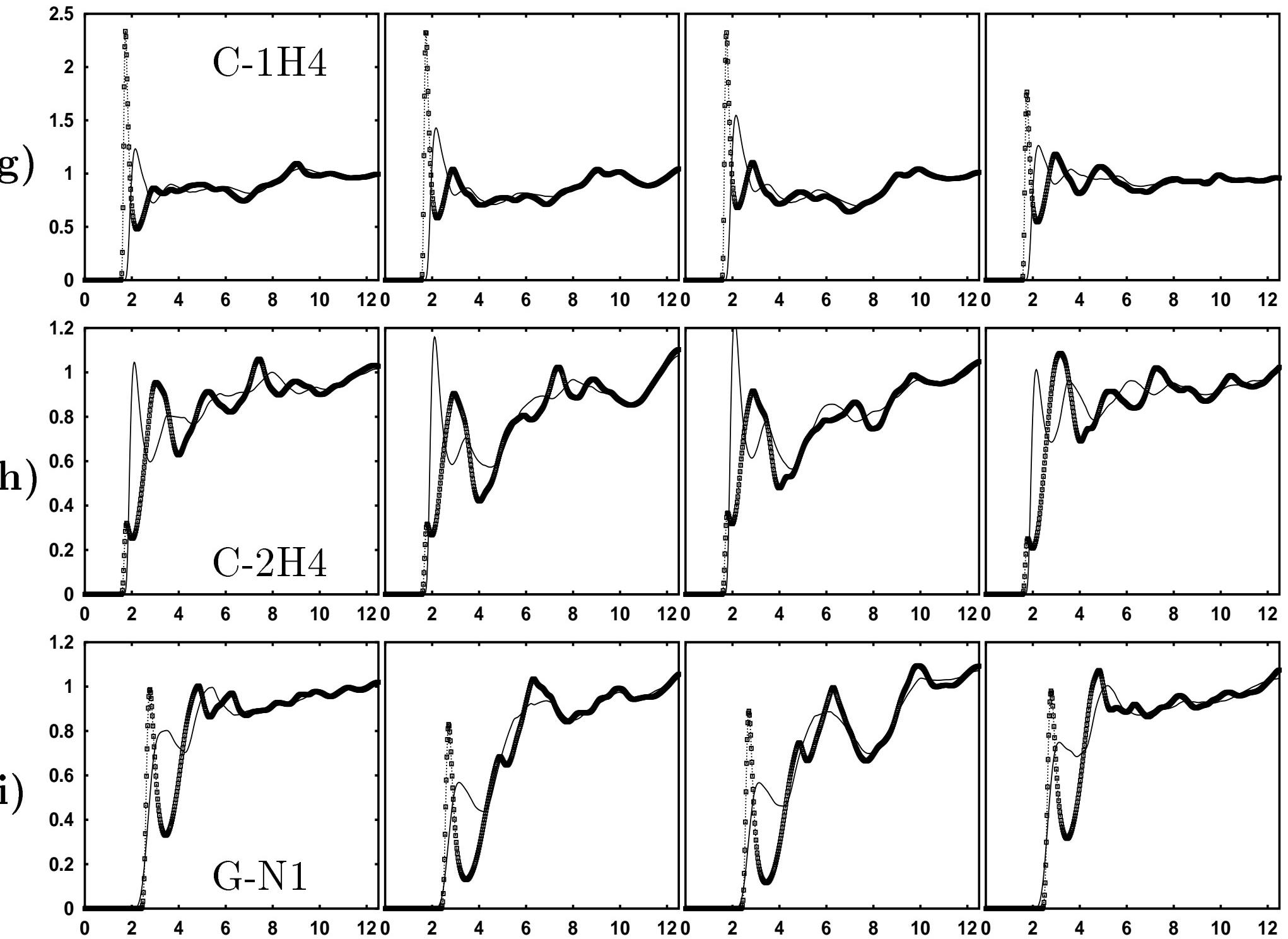
TABLE I. Computational expense for calculation of the DNA hydration. Here $N^{G:C}$ is the number of nucleotide pairs, N is the number of atoms in the fragment, N^W is the number of distinct elements in the structure matrix \hat{W} , N^{TFQMR} is the number of iterations required for solving the linear system, N^{ITER} is the total number of iterations for the given DNA fragment and τ_{PPro} is the computational time in minutes for a Pentium PRO 200 MHz and 64 Mb RAM computer.

$N^{G:C}$	N	N^W	N^{TFQMR}	N^{ITER}	τ_{PPro}
1	63	1038	124	7	12
2	126	1574	142	8	35
3	189	1715	149	8	70
4	252	1798	140	7	120
5	315	2020	148	7	190

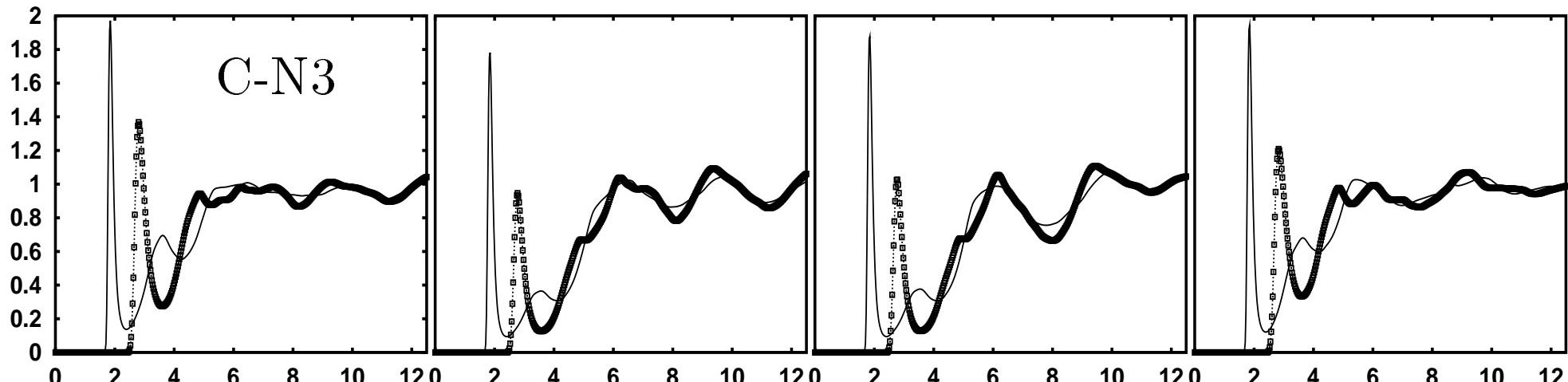
FIG. 1. Plots of the correlation functions, $h(r) + 1$, of some atoms in the DNA duplex with atoms of water versus the radius, r , measured in Angstöms. In each figure the solid line corresponds to the correlation function of the atom with the hydrogen atom of water molecules and the dashed line with points corresponds to the correlation function of the atom with the oxygen atom of water molecules. In each row we present from left to right the correlation functions corresponding to the given atom in the first, second, third and fifth nucleotide pairs of the B-DNA duplex d(GGGGG)-d(CCCCC) starting from the 5' atom in guanine. The data for the forth pair is suppressed.



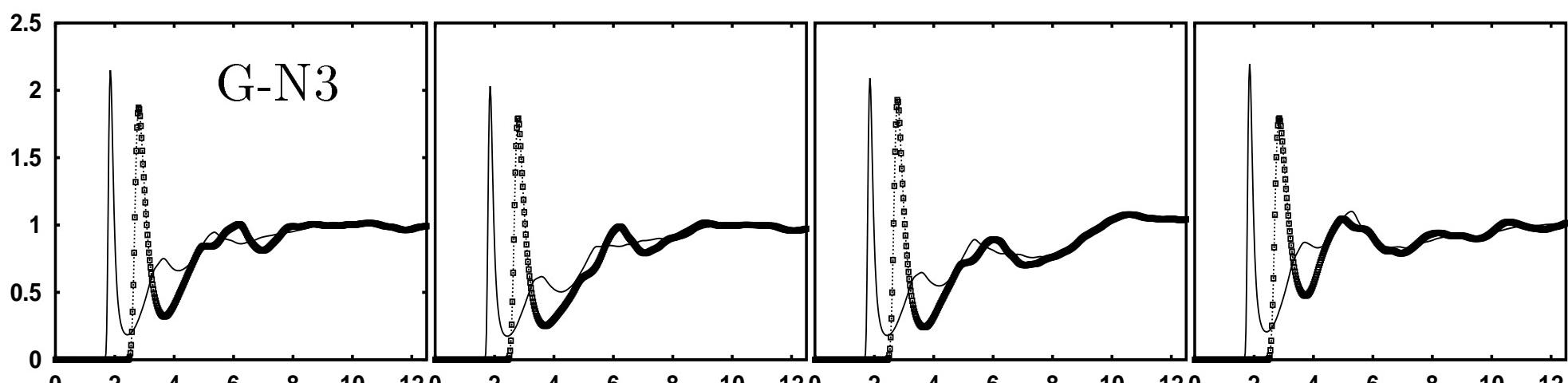




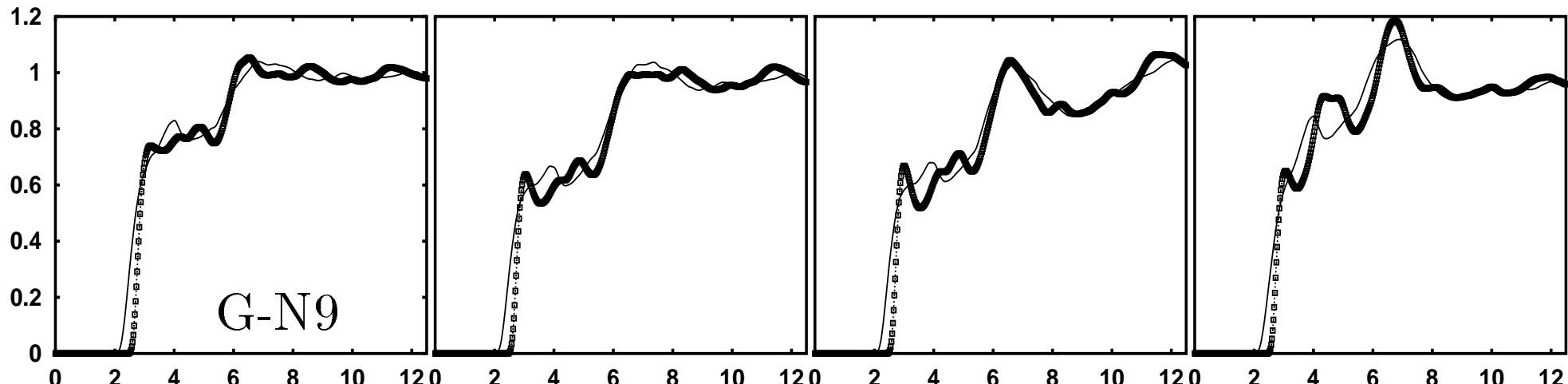
j)

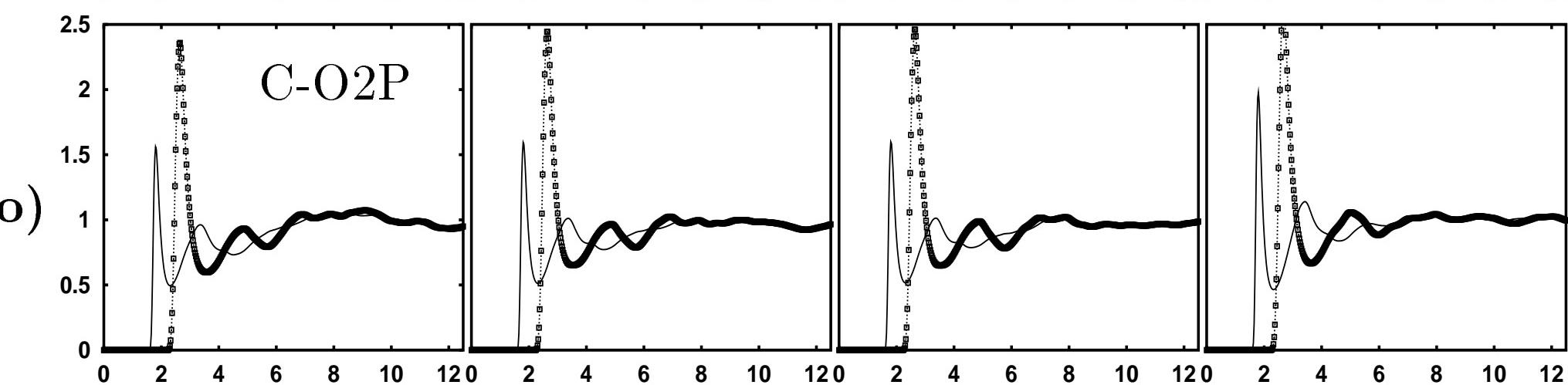
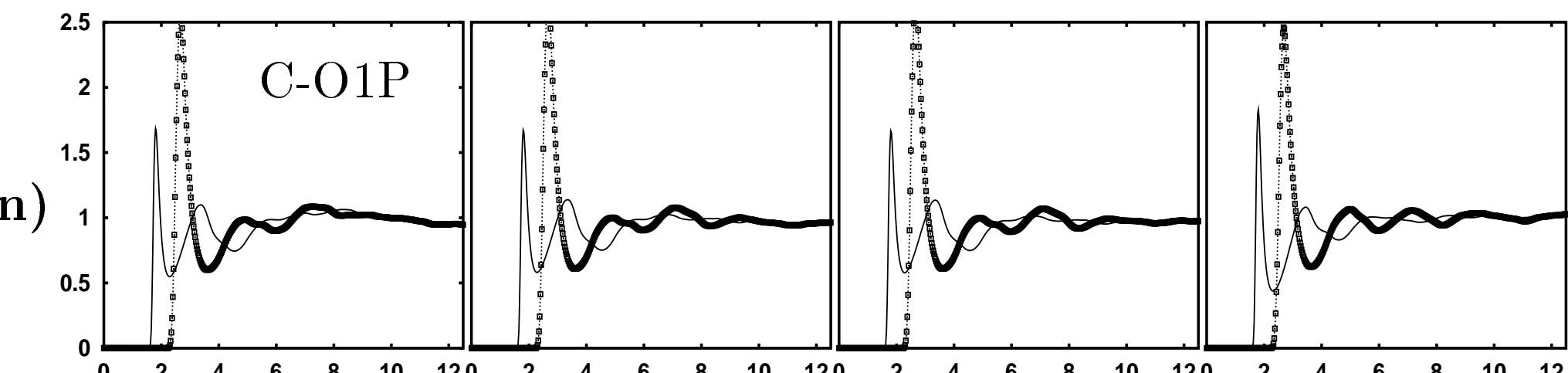
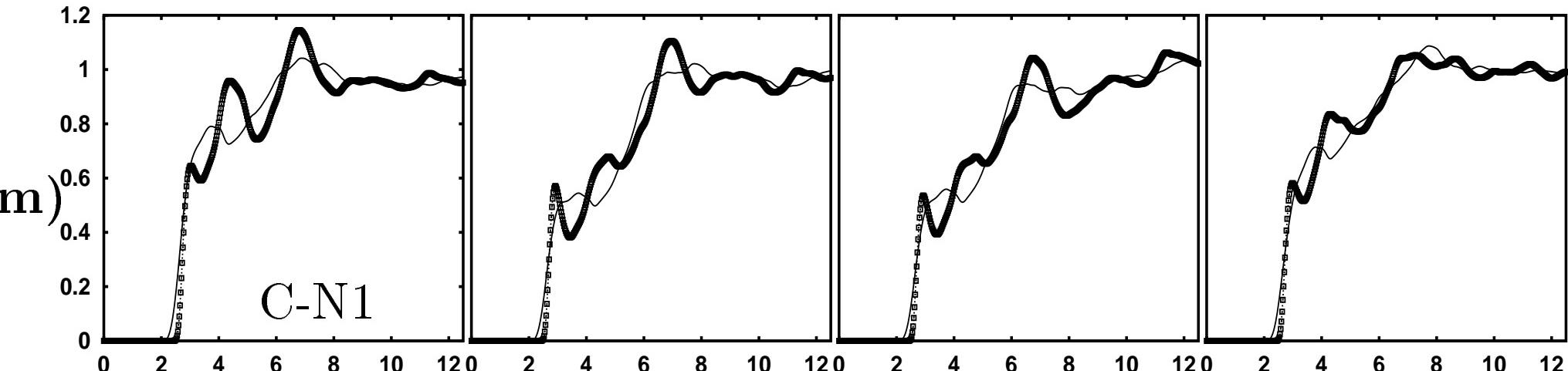


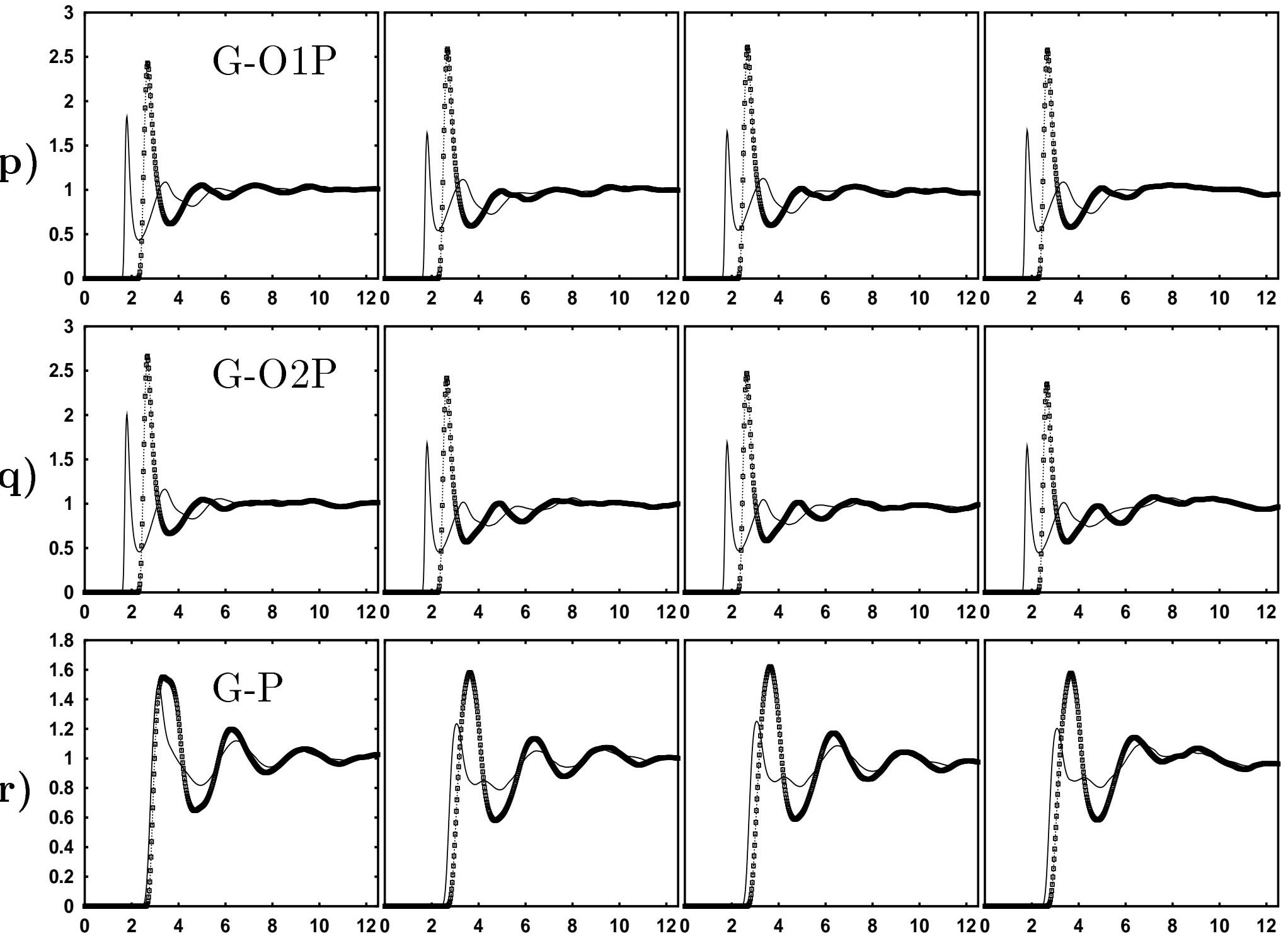
k)

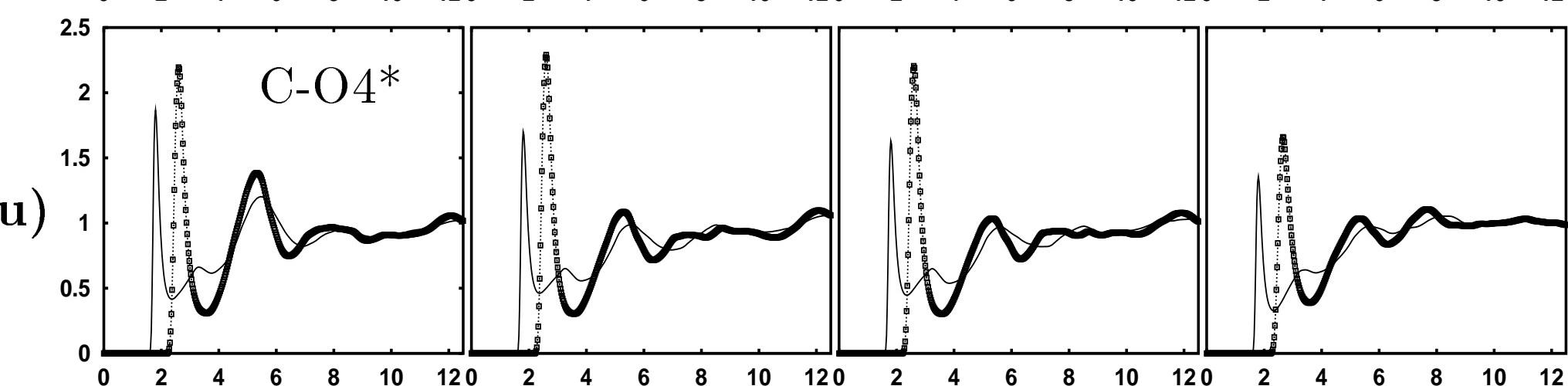
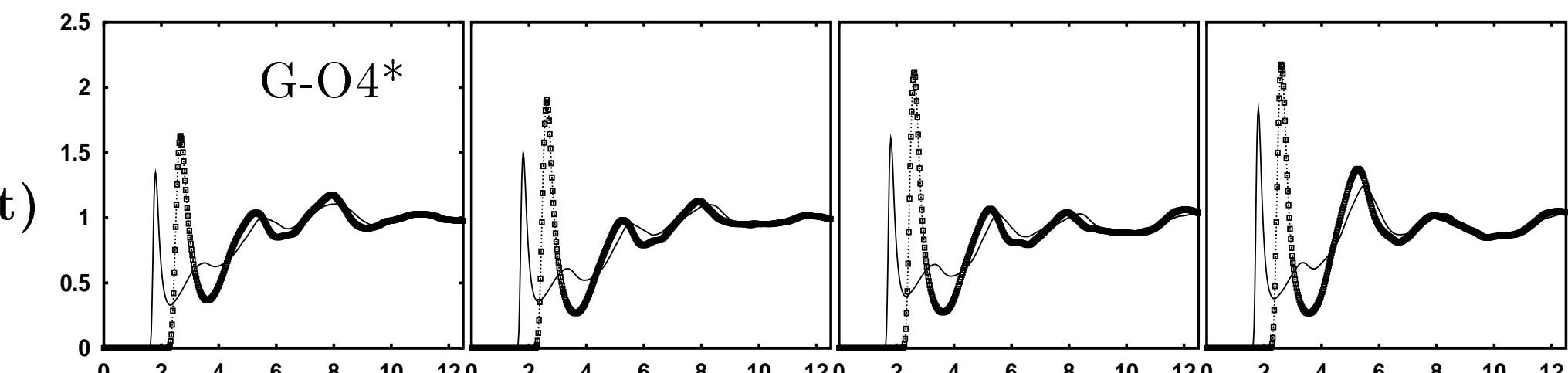
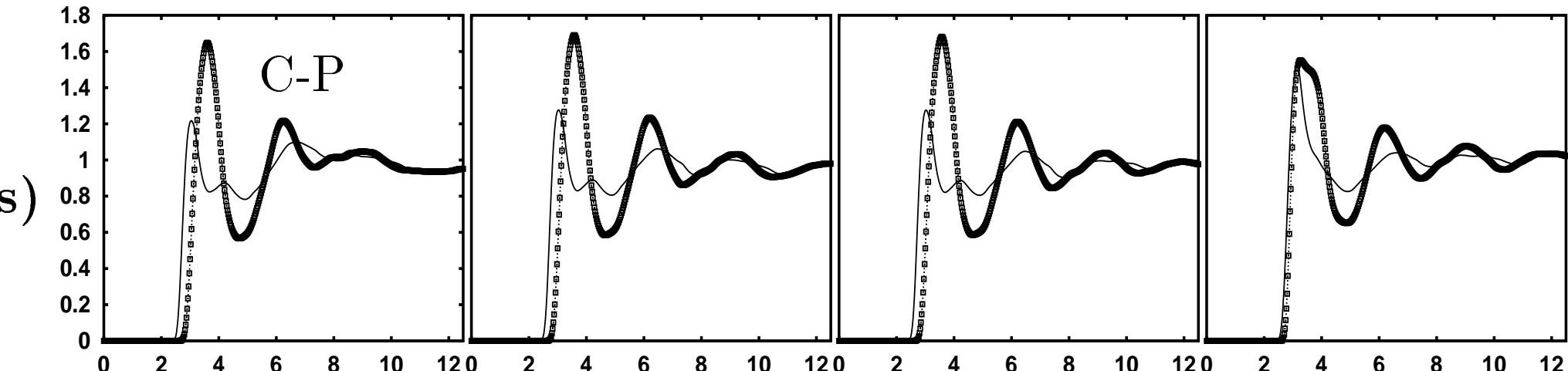


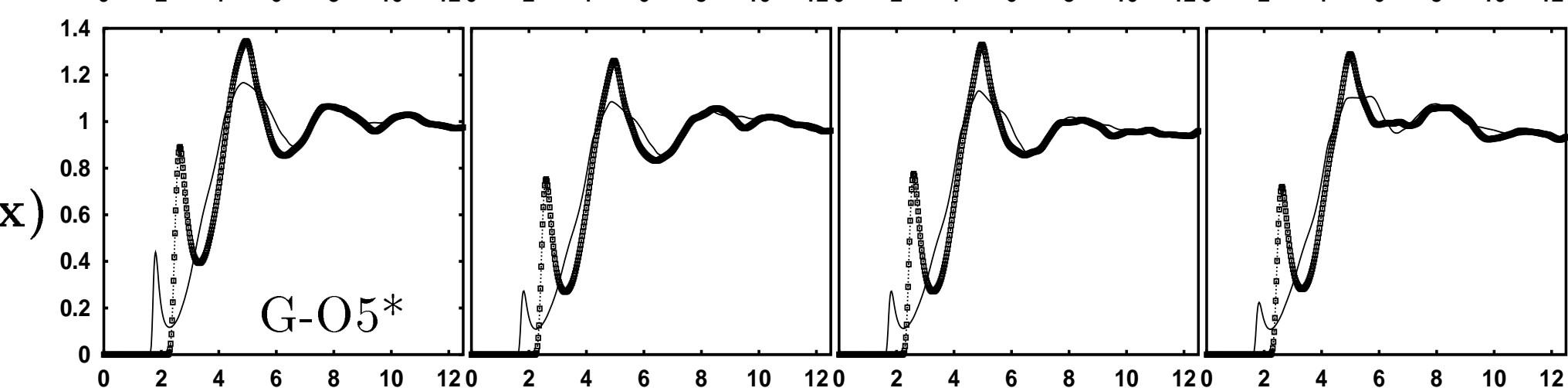
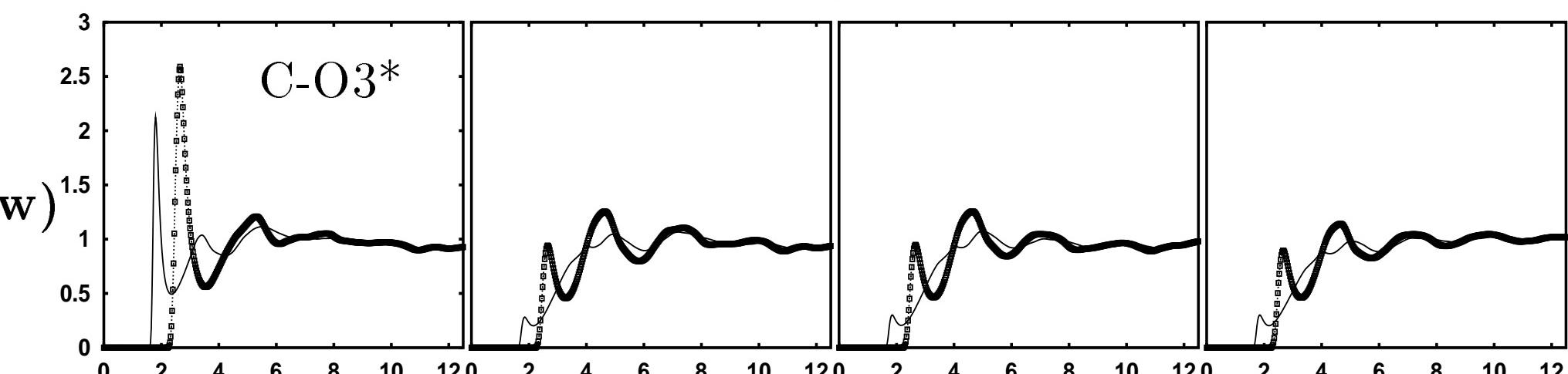
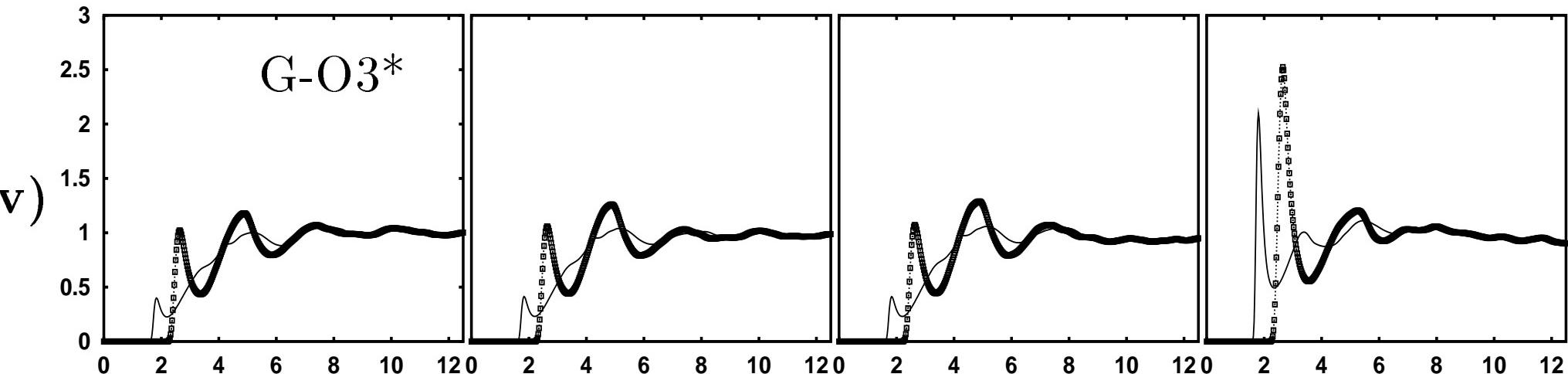
l)

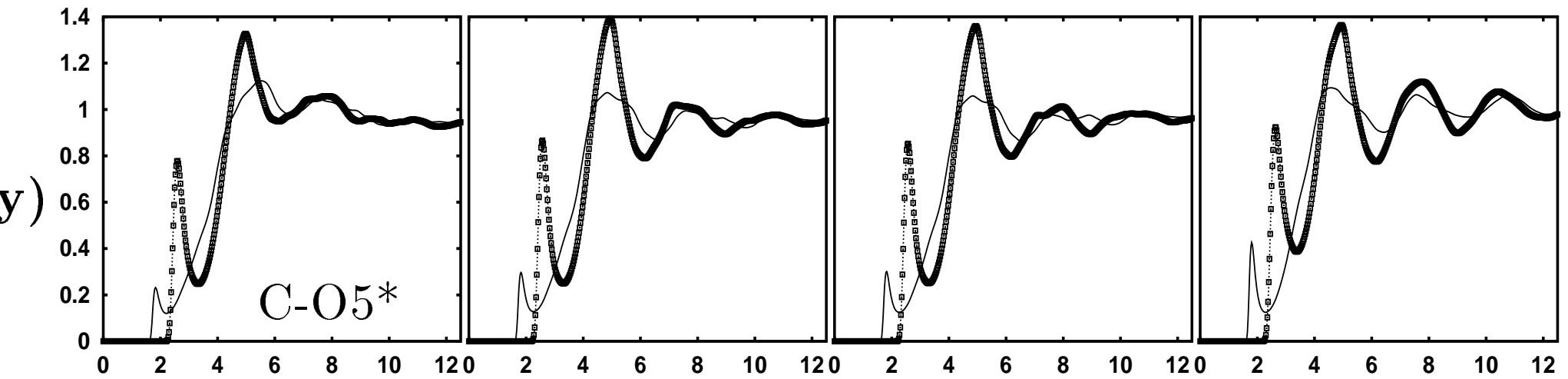




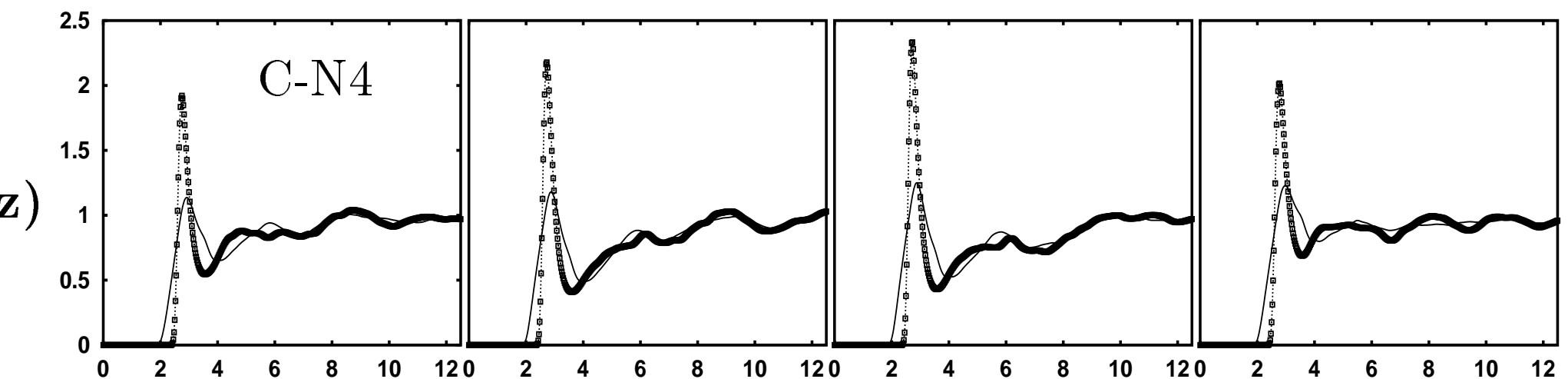








$C-O5^*$



$C-N4$